AD______________

Award Number: DAMD17-00-1-0119

TITLE: Interdisciplinary Breast Cancer Training Program

PRINCIPAL INVESTIGATOR: Coral A. Lamartiniere, Ph.D.

CONTRACTING ORGANIZATION: University of Alabama
Birmingham, AL 35294-0111

REPORT DATE: September 2005

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 01-09-2005	**2. REPORT TYPE** Annual Summary	**3. DATES COVERED** *(From - To)* 1 Sep 2004 – 31 Aug 2005

4. TITLE AND SUBTITLE
Interdisciplinary Breast Cancer Training Program

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-02-1-0318

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Coral A. Lamartiniere, Ph.D.

E-mail: coral@uab.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Alabama
Birmingham, AL 35294-0111

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The goal of the University of Alabama at Birmingham Interdisciplinary Breast Cancer Training Program (IBCTP) is to educate and train predoctoral students in a multidisciplinary environment with a focus on breast cancer research. The aims are to 1) recruit predoctoral trainees to the Interdisciplinary Breast Cancer Training program; 2) assure that predoctoral trainees obtain a broad-based breast cancer education and carry out interdisciplinary breast cancer research; 3) administer this program with sufficient oversight to ensure high-quality education and training, efficient completion of degree requirements, and productive research careers. We presently have 9 students that have either graduated or are academically in good standing. One has graduated and is doing postdoctoral breast cancer research at Duke University. Five students are accepted into Ph.D. candidacy, three others are into the 3rd, 2nd and 1st years. Of those nine, 2 are minorities. The IBCTP has hosted an active seminar program on cancer related research and provided the opportunity of the predoctoral trainees to talk to invited speakers. The Breast Cancer Causation and Regulation course and Breast Cancer Journal Club have received "very good" evaluations. Students have 2 manuscripts accepted and 6 submitted for publication. 19 Abstracts/presentations were made by students at scientific meetings. The students have received 13 awards, including one DOD predoctoral training grant and one Susan Komen predoctoral grant. Five new research grants were awarded to faculty, in part, because of student data used in the preparation of the grant applications.

15. SUBJECT TERMS
Breast cancer, interdisciplinary, predoctoral, training

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	24	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

INTRODUCTION

The goal of the University of Alabama at Birmingham Interdisciplinary Breast Cancer Training Program (IBCTP) is to educate and train predoctoral students in a multidisciplinary environment with a focus on breast cancer research. The aims are to 1) recruit predoctoral trainees to the IBCTP; 2) assure that predoctoral trainees obtain a broad-based breast cancer education and carry out interdisciplinary breast cancer research; 3) administer this program with sufficient oversight to ensure high-quality education and training, efficient completion of degree requirements, and productive research careers. Our training program is designed to prepare and motivate trainees to pursue careers in the fields of breast cancer causation, prevention, diagnosis, therapy and education.

BODY

The executive committee consist of: Dr. Danny Welch (Mechanisms of Growth Control), Dr. Therese Strong (Gene Therapy), Robert B. Diasio (Cancer Pharmacology), Clinton Grubbs (Chemoprevention), Charles N. Falany (Cancer Causation), and Dr. Coral A. Lamartiniere (Program Director), plus one elected student/trainee, Tim Whitsett. The executive committee is responsible for interviewing and selecting prospective IBCTP students, developing and implementing the academic and research program, review of individual student progress, the budget, and participating in Quarterly and Annual Program reviews.

TASKS FOR YEAR FIVE (No Cost Extension 9/04 - 8/05)

1) Schedule IBCTP seminar speakers (Aim 2).
The APPENDIX contains the list of breast cancer seminar speakers for 04 – 05. (pages 13 and 14).

2) Hold quarterly program reviews (Aim 3).
Quarterly program reviews were held by the executive committee to discuss recruitment, the progress of the trainees, the curriculum and the evaluation of courses. One new student was recruited: Scharri Ezell.

3) Monitor progress of trainees (Aim 3).
At the quarterly meetings, progress of individual students was discussed. At the end of the summer meeting, laboratory evaluations turned in by the mentors were taken into consideration. One of last year's first year students made satisfactory progress academically (A & B grades) and has selected a research mentor: Sarah Jenkins with Dr. Coral Lamartiniere (Cancer Causation and Regulation). Heath McCorkle dropped out of the program because of health problems. His fiancée died in an auto academic and Heath has suffered severe depression. He is under doctors' care. A list of students, research topic and mentors is provided in the APPENDIX (page 12).

4) Scientific Meetings and Abstracts (for all years)
Craig Rowell and Hope Amm attended and made poster presentations at the 2005 AACR meeting in San Francisco.

Craig Rowell attended and presented at 2005 Biostatics and Cancer meeting in Auburn AL.

Tim Whitsett attended and presented at the 2005 Gordon Conference on Hormone Action in Development and Cancer, and the 2005 Society of Toxicology Meeting in New Orleans.

James Cody attended and presented at the 2005 American Society of Gene Therapy Meeting in St. Louis, and at UAB's 2005 Student Research Day.

April Adams attended the 2004 AACR Special Conference on Chromatin, Chromosomes and Cancer Epigenetics, in Waikoloa, Hawaii.

Sarah Jenkins attended the 2005 7th International Symposium on Mass Spectrometry in the Health and Life Sciences (San Francisco, CA), the 2005 Breast Cancer and Environment Center meeting in Princeton and 2005 Clinical Proteomics Workshop: Today & Tomorrow (Nashville, TN - Vanderbilt University).

The PI attended and presented at the 2005 AACR meeting in San Francisco, 2 Breast Cancer and the Environment meetings in Cincinnati (2004) and Princeton (2005), and the 2005 Society of Toxicology meeting.

A list of student abstracts/presentations is contained in Reportable Outcomes.

5) Hold annual program review (Aim 3).
At the end of the summer executive committee meeting, the following recommendations were made. The Breast Cancer Causation and Regulation course and new format Breast Cancer Seminar Series received very good evaluations and it was recommended that the contents be kept the same. A copy of the Breast Cancer Causation and Regulation course content is enclosed in the APPENDIX (page 15).

6) Prepare and submit final report to DOD. Submitted.

KEY ACCOMPLISHMENTS

- The program now has 9 predoctoral Breast Cancer students in good academic standing and/or making good progress in breast cancer research or graduated.

- One student (Craig Rowell) has completed the requirements for his Ph.D. and has started a postdoc at Duke University with a continued research focus in breast cancer. Five (Damon Bowe, Tim Whitsett, James Cody, April Adams, Kevin Roarty) have been accepted into Ph.D. candidacy. One (Hope Amm) has an approved Ph.D. committee and is scheduling her qualifying exam. Sarah Jenkins has identified her mentor (Dr. Lamartiniere) and has started her research. Scharri Ezell is a minority first year student taking class work and carrying out lab rotations.

- For academic year 2004-2005, with only carry over funds, we interviewed 2 applicants (from 20 completed applications) and one was offered. Ms. Scharri Ezell, a miniority student was accepted. Her stipend and tuition is being paid from a UAB miniority fellowship.

- The appendix contains the lectures for the Breast Cancer Causation and Regulation course (page 12). The 2004-2005 course received a "very good" evaluation.

REPORTABLE OUTCOMES

- **Publications (for all years)**

Whisenhunt, T.W., Yang, X., **Bowe, D.B.**, Paterson, A.J., Toleman, C.A., Kudlow, J.E. "Escaping Repression at Estrogen Promoters: Regulated Coactivators in Repression Complexes." EMBO, *in review.*

Bowe, D.B., Yang, X.*, Muhkerjee, S., Whisenhunt, Rustgi, A.K., Paterson, A.P., Kudlow, J.E.: "Groucho/TLEs Repress Wnt Signalling Via *O*-GlcNAc Transferase." Nature Cell Biology, *in submission.*

Bowe, D.B., Adereth Y., and Maroulakou, I.G.: "ErbB2/Her-2 *neu* promotes mammary oncogenesis via reduction of $p27^{kip1}$ levels in cyclin D1-independent manner." Oncogene, *in submission.*

Sadlonova, A., Gault, S.R., Dumas, N.A., **Bowe, D.B.**, Van Tine, B.A., Mukherjee, S., Novak, L, Frost, A.R.: "Persistence and Growth-Inhibitory Effect of Human Breast Fibroblasts on the MCF10AT Xenograft Model of Proliferative Breast Disease." Cancer Research, *in submission.*

Bowe, D.B., Sadlonova A., Toleman, C.A., Hu, Y., Paterson, A.J., Kudlow, J.E.: "O-GlcNAc is a critical regulator of nuclear hormone receptor expression in mammary gland development." Molecular Cell Biology, *in submission.*

Bowe, D.B., Sadonlova A., Whiteside, M., Frost, A.R., Grizzle, W.E.: "CWR22 as a model for androgen sensitivity and androgen resistance of prostate cancer." Review article. (In preparation.)

Rowell, C., M. Carpenter, C. A. Lamartiniere, "Modeling Biological Variability in 2-D gel Proteomic Carcinogenesis Experiments" J. Proteome Res.; 2005; ASAP Web Release Date: 13-Aug-2005; (Article) DOI: 10.1021/pr0501261

Rowell, C., D. Mark Carpenter and Coral A. Lamartiniere. "Chemoprevention of Breast Cancer, Proteomic Discovery of Genistein Action in the Rat Mammary Gland." Accepted in Journal of Nutrition

- **Abstracts (for all years)**

Hope M. Amm, Patsy G. Oliver, Donald J. Buchsbuam. TRA-8 anti-DR5 antibody and chemotherapy agents produce cytotoxicity and activate apoptotic pathways in breast cancer cells. (Abstract #5357).

Bowe, D.B., Jones, M., Page, G.P., Allison, D.B., and Frost, A.R.: "Differences in gene expression of breast carcinomas of pre- and post-menopausal women." Era of Hope DOD Breast Cancer Research Program Meeting, Orlando, FL, Sept. 25-28, 2002.

Bowe, D.B., Jones, M., Sadlonova, A., Page, G.P., Allison, D.B., and Frost, A.R.: "Age-related gene expression profiles for invasive breast carcinomas in pre- and post-menopausal women." Mammary Gland Biology, Gordon Research Conference, Bristol, RI, June 1-6, 2003.

Whisenhunt, T.W., Yang, X., **Bowe, D.B.**, Toleman, C.A., Paterson, A.J., Kudlow, J.E. "Escaping Repression at Estrogen Promoters: Regulated Coactivators in Repression Complexes." Cambridge, U.K., March 18-21, 2004.

Cody, J., Lyons, G., and Douglas, J. A Dual-Action Armed Replicating Adenovirus for the Treatment of Bone Metastases of Breast Cancer. *Mol. Ther.* 9, S370, 2004.

Roarty, K and Rosa Serra. Wnt5a Exhibits a Growth Inhibitory Effect on Development of the Mammary Gland, American Society for Cell Biology 45th Annual Meeting, San Francisco, CA 2005.

Rowell, C, Isbell, S, Desilva, T and Lamartiniere, CA. 2-Dimensional gel electrophoresis and proteomic identification of mammary gland proteins of rats treated with the soy isoflavone, genistein. Proceedings of the American Association for Cancer Research. 43:35, 2002.

Rowell, C., Whitsett, T., Carpenter, M. and Lamartiniere, C.A. Proteomic Analysis of Uterine Proteins Following Genistein Exposure. Proceedings of the American Association for Cancer Research. 44: 713, 2003.

Carpenter, M., **Rowell, C,** Lamartiniere, C. and McCorkle, H., "2D-gel Proteomics in biomarker discovery." In Proceedings of Pharmaceutical Industry SAS Users Group 2004, San Diego, California.

Rowell, C., C. Lamartiniere, "Discovery of a Novel Pathway of Chemoprevention by Genistein using Proteomics" Susan G. Komen Mission Conference, New York, NY, 2004.

Rowell, C., G. Puckett, K. Roarty, M. Kirk, L. Wilson, M. Carpenter and C. A. Lamartiniere, "Serum profiling and biomarker discover of rat mammary tumors using mass-coded abundance tags (MCAT)" 95th Annual meeting of the American Association for Cancer Research, Orlando, FL, 2004.

Rowell, C. and C.A. Lamartiniere. From Discovery to Validation: Statistical and Biological evaluations of Proteomics data. Department of Mathematics and Statistics, Auburn University, Auburn, AL 2005

Rowell, C. and C.A. Lamartiniere. Proteomic Discovery of Genistein Action in the Rat Mammary Gland. Craig Rowell and Coral A. Lamartiniere, 2005 AACR meeting in San Francisco.

Whitsett, T. and Lamartiniere, C.A. Genistein regulates GRIP-1 in the rat mammary and uterus. Presented at South Central Society of Toxicology Meeting in Chattanooga TN, September, 2003.

Whitsett T, Wang J, and Lamartiniere CA. Steroid coactivator GRIP-1 regulation with genistein in the rat mammary gland. AACR Annual Meeting. Proceedings, Volume 45:661. 2004.

Whitsett T, Wang J, and Lamartiniere CA. Genistein regulates the steroid coactivator GRIP-1 in the rat mammary gland. Society of Toxicology 43 Annual Meeting. Program page 58. 2004.

Whitsett T and Lamartiniere CA. Breast Cancer Chemoprevention with the Polyphenol Resveratrol. Emerging Topics in Breast Cancer and the Environment Research. 2004.

Whitsett T and Lamartiniere CA. Breast Cancer Chemoprevention with the Polyphenol Resveratrol. Gordon Research Conference: Hormone Action in Development and Cancer. July 2005.

Whitsett T and Lamartiniere CA. Breast cancer chemoprevention with the polyphenol resveratrol. Society of Toxicology 44 Annual Meeting. Program page 164. March 2005.

3) Awards to Predoctoral Students (for all years)

April Adams: AACR Minority Travel Scholar Award in Cancer Research, November 2004

Damon Bowe: Merck Toxicology Externship, Safety Assessment Division, Merck & Co., West Point, PA, May 2005

James Cody: Elected Presiding Officer in the Molecular and Cellular Pathology graduate program for both the '04-'05 and '05-'06 academic years.

Craig Rowell: Susan Komen Breast Cancer Predoctoral Award (DISS0201242) Effects of Genistein and TCDD on the Maturation of the Rat Mammary Gland: Alterations in Protein Tyrosine Kinase Activity and Signaling.

Craig Rowell: "AACR Scholar in Training Award" Travel award for the 2004 AACR meeting

Craig Rowell: 1st place for scientific posters sponsored by the Breast Cancer and the Environment Research Centers (BCERC) in November 2004 in Princeton NJ

Craig Rowell: Awarded Graduate Student of the Year, Department of Pharmacology and Toxicology, 2005

Tim Whitsett: Southeastern Society of Toxicology Poster Award (2003)

Tim Whitsett: 2nd plasce for Emerging Topics in Breast Cancer and the Environment Research Poster Award (2004)

Tim Whitsett: Susan G. Komen Foundation Travel Scholarship (2004)

Tim Whitsett: Graduate Student-Postdoctoral Fellow Conference Award (Gordon Research Conference 2005)

Tim Whitsett: Society of Toxicology Travel Award (2005)

Tim Whitsett: DOD Predoctoral Training Award (BC043793) Chemoprevention Against Breast Cancer with Genistein and Resveratrol. 2/25/05 - 2/25/08

4) Research grants received in part because of preliminary data produced by Breast Cancer predoctoral students (for all years)

NIEHS 1R21 ES012326-01 (C.A. Lamartiniere, PI) 4/18/03 – 3/30/06
First Year: $100,000; Total: $300,000
In Utero TCDD Programming for Mammary Cancer: Proteomic analysis of mammary gland from rats treated in utero with TCDD.

DOD DAMD BC 17-03-1-0433 (C.A. Lamartiniere, PI) 7/1/03-7/31/06
First Year: $150,000; Total: $428,249
Proteomic Analysis of Genistein Mammary Cancer Chemoprevention: Proteomic analysis and interstitial fluid analysis of mammary glands of rats treated with genistein.

Center for Nutrient-Gene Interaction in Cancer Prevention. NIH NCI P20 CA93753-02, S. Barnes, Center Director. Project 1. Polyphenols: Mammary and Prostate Cancer Chemoprevention. (C.A. Lamartiniere, C.A., P.I.). $833,638. 6/1/03-9/30/08.

Center for the Study of Environment and Mammary Gland Development. NIH/NIEHS. 1U01 ES012771-01. J. Russo, Fox Chase Cancer Center, Director; Lamartiniere, Co-PI. 9/29/03 – 7-31-/10. UAB PI share: $1,540,000.

NIH 1 R01 CA108585-01A2 , Armed Replicating Ad for Breast Cancer Bone Metastasis, Joanne Douglas, PI.

Summary. The UAB institutional predoctoral breast cancer training grant has been a success on this campus. It has catered to a subset of focused bright young students/researchers that are dedicated to investigating the cause, chemoprevention and therapy of breast cancer. These young researchers are being trained to carryout cutting edge breast cancer research. While the BCTP has been in existence for only 5 years, we have graduated one Ph.D. who is carrying out breast cancer research at Duke University. Another is expected to graduate with his Ph.D. this year. Then, we expect 6 more to graduate within the following 2 years. Overall we expect 9 Ph.D.s in breast cancer research, 2 who are minorities. We are optimistic about the productivity of these students based on the short term published and submitted manuscripts and the abstracts presented at national/international meetings. Productivity will be better measured in the coming 5 years.

UAB is appreciative of the opportunity of hosting a DOD breast cancer training program.

APPENDIX

Student Credentials

Student Research and Mentors

IBCTP Seminar Speakers

2004-2005 Breast Cancer Causation and Regulation Lectures

One Manuscript In Press

Students Who Matriculated in the University of Alabama at Birmingham Interdisciplinary Breast Cancer Training Program

	Previous Degree	Date of		GRE			
Student	Institution	Entry	GPA	Verbal	Quantitative	Analytical	Status
Craig Rowell	BS (95) Lake Forest IL MS (00) UAB	2000	3.8	580	610	680	Postdoc Duke U
Chantelle Bennetto	BS (99) U. Saskatoon	2000	4.0	510	660	710	Switched to Pharmacology
Mubina Nasrin	MD (94) M.R. Medical College, India	2001	no GPA	690	650	670	Med School UAB
Damon Bowe	BS (99) Bates College Maine	2001	3.5	590	580	710	Accepted into Ph.D. Candidacy
Hope Amm	BS (02) Saint Mary's College	2002	3.38	550	640	490	Ph.D. committee has been approved
Timothy Whitsett	BS (02) Yale University	2002	3.59	530	700	750	Accepted into Ph.D. Candidacy
James Cody	BS (01) UAB	2003	3.37	590	670	640	Accepted into Ph.D. Candidacy
April Adams	BS (01) U. Chicago	2003	3.38	660	710	-	Accepted into Ph.D. Candidacy
Kevin Roarty	BS (95) Virg. Tech. M.S. (02) UAB	2003	3.74	520	680	480	Accepted into Ph.D. Candidacy
Sarah Jenkins	BS (04) St College of West Georgia	2004	3.5	550	690	-	Into second year research and academics
Heath McCorkle	BS (01) Emory Univ	2004	3.1	570	740	-	Withdrew because of health problems
Scharri Ezell	BS (05) Tougaloo College	2005	3.5	480	710	-	Matriculated

Students, Breast Cancer Research, and Mentors

Student	Reseach Descrption	Mentor (Department)
Craig Rowell	Proteomic Discovery of Genistein Action in the Rat Mammary Gland	Coral Lamartiniere (Pharmacology and Toxicology)
Mubina Nasrin	Therapeutic Potential of UAB 30 (a retinoic acid derivative) against Breast Cancer	Coral Lamartiniere (Pharmacology and Toxicology)
Damon Bowe	NCOAT Splice Variant Function in Mammary Gland Development	Jeffrey Kudlow (Molecular Endocrinology)
Hope Amm	Combination Immunotherapy and radiation Therapy against Breast Cancer	Donald Buchsbaum (Radiation Oncology)
Timothy Whitsett	Chemoprevention Against Breast Cancer with Genistein and Resveratrol	Coral Lamartiniere (Pharmacology and Toxicology)
James Cody	Gene Therapy for Breast Cancer	Joanne Douglas (Pathology)
April Adams	Molecular Imaging in Animal Models	Kurt Zinn (Radiation Oncology)
Kevin Roarty	Mechanisms of TGF-beta Action in Mammary Gland Development and Breast Cancer	Rosa Serra (Cell and Molecular Biology)
Sarah Jenkins	Endocrine Disruptors and Breast Cancer	Coral Lamartiniere (Pharmacology and Toxicology)
Scharri Ezell	Breast cancer therapy	Robert Diaiso (Pharmacology and Toxicology)

Breast Cancer Training Program Seminars

2004 - 2005

October 5, 2004
Graeme Bolger, M.D., Associate Professor of Medicine, Med-Hematology & Oncology, UAB
"Regulation of cAMP Signaling Pathways"

October 12, 2004
Sue Heffelfinger, Ph.D., Associate Professor, University of Cincinnati Department of Pathology & Laboratory Medicine
"Angiogenesis: A regulator of Mammary Tumorigenesis"

October 19, 2004
John Hartman/ Genetics
"Genetic Buffering of Ribonuncleotide Reductase"
Genetics, UAB

October 26, 2004
Sandra Haslam. Ph.D., Professor and Director, Breast Cancer and Environmental Research Center, Michigan State Univ
"Progesterone Action in Normal Mammary Gland Development and Breast Cancer"

November 2, 2004
Hitoshi Someya, Graduate Student, Pharmacology and Toxicology, UAB
"Mechanism of Action of 4'-thio arabinofuranosylcytosine (Tara C)"

November 16, 2004
Martin Johnson, Ph.D., Professor of Pharmacology & Toxicology, UAB. "Rationally Designed Treatment for Cancer: Is it really Rational?"

December 7, 2004
Zhiyuan Shen, Ph.D, Associate Professor, Department of Molecular Genetics and Microbiology, University of New Mexico School of Medicine
"Protection of Genomic Integrity by a BRCA2 Interacting protein: BCCIP"

January 11, 2005
Catherine Chaudhuri, Ph.D., Professor of Chemistry and Biochemistry, University of Maryland
"Organelle proteomics to study acquired drug Resistance"

January 21, 2005
Xianglin Shin, Ph.D., Professor of Microbiology, Immunology, and Cell Biology, Department of Genetics and Developmental Biology, West Virginia University
"Antioxidant Properties of Apple Peel Extract and Tumor

	Prevention"
February 8, 2005	Chantelle Bennetto, UAB "Novel Antiretroviral Quantitation Methodologies"
March 1, 2005	Carlos Sonnenschein, M.D., Tufts University "The Tissue Organization Field Theory of Carcinogensis: New Perspectives"
March 8, 2005	Amanda Foxwell, Graduate Student, Pharmacology and Toxicology, UAB "Structural and Functional Effects of Aldehyde Modification of Mitochondrial Proteins"
March 15, 2005	Marilyn Moore, Ph.D., "Flavonoid-Drug Interactions: Effects of flavonoids on ABC" University of Buffalo
April 21, 2005	Craig Rowell, Pharmacology and Toxicology Graduate Student, UAB, Dissertation Defense, Candidate for the Degree of Ph.D. in Pharmacology & Toxicology, "Discovery Proteomics: Model Development and Validation in the Rat Mammary Gland"
April 26, 2005	Tracy D'Alessandro, UAB Department of Pharmacology & Toxicology, "Soy isoflavones: complex metabolism of an antioxidant class"
May 3rd, 2005	Gary Piazza, Ph.D., Adjunct Associate Professor, Southern Research Institute, "Soy isoflavones: complex metabolism of an antioxidant class"

Breast Cancer Causation and Regulation
TOX 750
Spring/Summer 2005
Mondays and Wednesdays, 3-5 pm in Volker Hall 108D
Course Director: Coral A. Lamartiniere
Volker Hall 124; 4-7139; Coral@uab.edu
Administrative Coordinator: Sharon Bohannon Volker Hall 108H; 4-4579; sbohannon@ccc.uab.edu

Date	Topic	Instructor (Department)
Mon April 4	Overview of the Breast Cancer Problem	John Waterbor (Epi)
Thur April 14	3:00 pm Environmental Carcinogenesis	Coral Lamartiniere (Pharm/Tox)
Mon April 11	Steroid Hormone Action in the Breast	Barnes (Pharm/Tox)
Wed April 13	Oncogenes and Suppressor Genes	Mike Ruppert (Medicine)
Mon April 18	Signal Transduction and Breast Cancer	Jeffrey Kudlow (Endocrinology)
Wed April 20	Exam	
Mon April 25	Nuclear Receptors as Targets for Novel Small Molecule Therapeutics	Donald Muccio (Chemistry)
Wed April 27	Cancer Pharmacology	Robert Diasio (Pharm/Tox)
Mon May 2	Cancer Metastasis (Mechanisms)	Danny Welch (Pathology)
Wed May 4	Chemically-induced Models of Breast Cancer (Chemoprevention)	Clinton Grubbs (Chemoprevention)
Mon May 9	Primary Prevention	Mona Fouad (Preventive Medicine)
Wed May 11	Exam	
Mon May 16	Breast Cancer Metastasis	Joanne Douglas (Pathology)
Wed May 18	Targeted Immunotherapy	Denise Shaw (Medicine)
Mon May 23	Tumor-host/stroma interactions	Rosa Serra (Cell Biol)
Wed May 25	Pathology of Breast Cancer	Andra Frost (Pathology)
Mon May 30	Gene Therapy	Theresa Strong (Gene Therapy)
Wed June 1	Exam	

research articles Journal of proteome research

Modeling Biological Variability in 2-D Gel Proteomic Carcinogenesis Experiments

Craig Rowell,[†] Mark Carpenter,[§] and Coral A. Lamartiniere*[,†,‡]

Department of Pharmacology and Toxicology, UAB Comprehensive Cancer, University of Alabama at Birmingham, Birmingham, Alabama 35294, and Department of Mathematics and Statistics, Auburn University, Auburn, Alabama

Received May 3, 2005

We propose a statistical method to model the underlying distribution of protein spot volumes in 2-D gels using a generalized model (GM). We apply this approach to discover mechanisms of chemical carcinogenesis in a rodent model. We generated 247 protein spots that were common to all gels (n = 18). Traditional statistical methods found 6.5% (13 out of 247) significant protein spots, our GM approach yielded a total of 53 (22.5%) differentially expressed protein spots.

Keywords: statistics • 2-D gels • proteomics • carcinogenesis • DMBA • rat

1.0. Introduction

Since the first major studies using two-dimensional gel electrophoresis (2-D gels), the field of proteomics has undergone rapid growth and development.[1] Coupled with mass-spectrometry based protein identification, 2-D gels have been viewed by scientists as a tool for the discovery of proteins and pathways in numerous systems.[2–5] Progress in proteomics research has been directly related to the availability of standard reagents, protocols, and computer programs for data analysis (i.e., Progenesis and PDQuest).[6] These improvements have increased the number of treatment/comparison groups as well as the number of biological replicates within each group that can be examined. In addition, better imaging and processing software allows for attention on proper statistical design and analysis of experiments.

Postrun analysis is the bottleneck of 2-D gel experiments due to high dimensional data likely having high variability.[7] Deficiencies in experimental design and execution greatly impede postrun analysis and decrease the overall sensitivity of the technique. Problems related to analysis first arise in the software processing of the gels, as reported in Nishihara and Champion.[8] These results point to the issue of false positive discovery vs accuracy as a tradeoff affecting the choice of software to use. Another consideration is building composite gels to increase the number of real spots to analyze. Central to composite gel analysis is how to treat absent spots (i.e., averaging intensities vs choosing a "best-of" analysis). Technology such as CyeDyes can potentially overcome this problem, but not without introducing other considerations. Mauer et al. examined 2-D gel data using statistical processes inherent in the analysis software as well as algorithms applied to microarray data.[9] Recently, Chang et al. investigated the issue of spot normalization (a computer generated process) to address the issue of missing values (spots that are represented in the majority but not all of the gels in a data set).[10] A modeling procedure used by Gustafsson et al. adjusts for variances in spot volume data by applying alternative transformations.[11] That each of the above approaches has had a measure of success shows there are numerous approaches for evaluation of 2-D gels.

Much of 2-D gel analysis is based on the search for significant variation between the means (medians) in different groups using the two-sample t-test and analysis of variance (ANOVA). The assumption is that the populations being studied are normally distributed with constant variances, independent of the mean expression levels. If the assumptions are violated, transformations (i.e., log) are taken to make the data more closely conform to the normal distribution. However, this approach has produced limited success because the transformation is usually taken across all analysis variables. Gustafsson et al. noted that even after they transformed their 2-D gel expression data, substantial variance heterogeneity remained.[11] So, rather than manipulating the data until it conforms to preconstructed assumptions, we propose to model the data separately for each protein.

From evolution and development literature we borrow the term "standard norms of reaction" (NoR) to introduce our modeling process of 2-D gel data. Woltereck introduced the concept of NoR to represent the variation of phenotypic response to environmental alterations based on the genotype of the organism.[12,13] The environmental condition in our current study is the process of carcinogenesis. In this study the same genetic strain of animals has been exposed to the same environmental insult (dimethylbenz[*a*]anthracene, DMBA). We know that this experiment will result in the production of mammary tumors in all treated animals. We also know that the timeline of palpable tumor development is variable among

[*] To whom correspondence should be addressed. E-mail: Coral@uab.ed.
[†] Department of Pharmacology and Toxicology, University of Alabama at Birmingham.
[‡] UAB Comprehensive Cancer, University of Alabama at Birmingham.
[§] Auburn University.

10.1021/pr0501261 CCC: $30.25

PAGE EST: 8.5

Published on Web 00/00/0000

the individual animals; therefore, there is an underlying plasticity in the phenotypic response.[14,15] To avoid confounding effects of tumor heterogeneity, we will look at the period of early lesion formation.[14] In general, we presume that changes observed at this time point will reflect early biochemical events related to promotion. It is our goal to model a tissue protein signature(s) associated with early cancer formation.

In this paper, we describe the importance of statistical design and analysis when conducting investigations using 2-D gels for differential protein expression profiling. A series of experiments and analyses related to our research into the biochemical mechanisms of carcinogenesis by DMBA in a rodent mammary model provide the data. We propose a statistical method whereby the underlying distribution of spot volume is modeled directly as a generalized distribution. This generalized model (GM) encapsulates the various methods of transformations and analyses found in modern proteomic literature. The GM method will therefore yield better rates of discovery than more traditional proteomic statistical analyses and better reflect biological changes in protein expression.

2.0. Materials and Methods

Sprague–Dawley CD rats were purchased from Charles River Breeding Laboratories (Raleigh, NC). Dimethylbenz[*a*]anthracene (DMBA) and sesame oil were purchased from Sigma Chemical Company (St. Louis, MO). Isoelectric focusing (IEF) strips, IEF buffer, Multiphor II, tissue grinding kits, and albumin removal kits were purchased from Amersham Biosciences (now a member of GE Healthcare, Piscataway, NJ). All other chemicals were purchased from Fisher Scientific (Hampton, NH). SyproRuby and the VersaDoc densitometer were purchased from Bio-Rad (Hercules, CA). SAS v.10 was purchased from the SAS Institute (Cary, NC).

2.1. Pilot Projects, Replication and Power Analysis. One important aspect of experimental design is choosing sample size.[16] In this study, we promote the use of power analysis in determining sample size. Other issues of statistical design are the elimination of extraneous sources of variability and choosing the number and levels of comparison groups. Our first consideration is the choice between technical and biological replications.

Technical versus Biological Replication. Using technical (analytical) replicates over biological replicates has been widely discouraged.[17-19] However, Asrivatham et al. stated that the use and investment in analytical replicates for pilot projects is extremely valuable for data quality control and validation of the 2-D gel handling process.[20] Importantly, the general consensus is that as functions of cost and resources, biological replicates provide considerably more scientific information than analytical replicates.

Power and Sample Size. Pilot studies were conducted to determine optimal sample size based on power analysis. In our study, the power estimate and sample size determinations involved using unique uterine samples (biological replicates) from 8 control- and 8 genistein- (a phytoestrogen found in soy) treated rats. The variance for each of the commonly detected proteins was estimated using the pilot expression data. The variance estimate was used to evaluate sample size effects for discovering specific protein volume fold-changes. Rather than basing power analysis on crude family wise adjustments, such as Bonferroni, we designed an experiment with sufficient power to examine at least one single protein comparison (in our study, the power analysis was based on adjustment of 100 proteins) and after the data was collected we computed the estimated false discovery rate to assess the potential number of discoveries.

False Discovery Rates (FDR). Benjamini and Hochberg first coined the phrase "false-discovery rate" (FDR) now commonly applied in significance testing designed for high dimensional biology.[21-25] For a particular experiment, the FDR is the expected or estimated proportion of false discoveries out of the total number of significantly different genes/proteins. This means that a large FDR of 50% would lead the researcher to a different decision with respect to allocation of resources than if the FDR were 5%. Therefore, we computed the estimated FDR to assess the potential number of discoveries after the data was collected.[23-24]

2.2. Study Design. Animal care and treatment were performed according to established guidelines approved by the UAB Animal Care Committee. Eighteen 50-day-old female Sprague–Dawley rats were divided into two groups and either gavaged with 40 μg DMBA/g B. W. ($n = 8$) or gavaged with an equal volume of vehicle, sesame oil only ($n = 10$). At 75 days of age (25 days post DMBA treatment), animals were anesthetized with Ketamine/xylazene and the fourth abdominal mammary glands were dissected. We selected 75 days post DMBA with the intention of investigating mammary glands with early preneoplastic lesions and biochemical alterations, and yet relatively tumor mass free. Each gland was cut in half longitudinally to allow both proteomic as well as pathological evaluations. Frozen mammary tissues were homogenized in lysis buffer formulated for 2-D gels using tissue grinding kits.[26] After measuring protein concentration via Bradford's assay (Bio-Rad), equal concentrations of sample were subjected to albumin removal. Protein concentration was remeasured and 150 μg protein aliquots were diluted in rehydration buffer. The samples were applied to separate immobilized pH gradient (IPG) strips (24 cm, pH 4–7) and allowed to rehydrate overnight at room temperature. The IPG strips were placed on a flatbed electrophoresis unit (Multiphor II) and a current gradient applied (500 V for 1 h, 3500 V for 1.5 h, followed by 3500 V for 22.5 h). After isoelectric focusing, IPG strips were equilibrated first in 100 mM dithiothreitol for 45 min followed by equilibration in 120 mM iodacetimide for 45 min. IPG strips were loaded onto pre-cast 1.5 mm, 12.5% SDS gels and run on a Dodecacell vertical electrophoresis unit according to manufacturer's suggestions. Both IEF and SDS gels were run as block groups consisting of equal treatments per run. Once gels were run to completion, they were stained using SyproRuby and scanned via a VersaDoc 4000 Densitometer. Spot matching and gel warping were done using Progenesis Discovery 2004. Processed data was imported into SAS version 10 and analyzed using statistical methods and algorithms based on various SAS procedures.

2.3. Data Processing. For our experiments, we elected to use the "total spot volume" normalization procedure found in the Progenesis software. After spot matching and gel warping were completed, the data file was exported to SAS for processing. For all the following procedures, we evaluated only spots that were common to all gels in the data set. The first step in our cleanup procedure was to perform a t-test. Each spot identified as significant ($p < 0.05$) was located and the spot's presence was visually confirmed in all the gels. As needed, manual rematching of spots was conducted and the statistical program was re-run to generate a new list of p-values for the matched spots. This iterative process was run numerous times to ensure that matches reflected high quality spots (i.e., consistent shape, nonsaturation and proper splitting).

2.4. Statistical Methods. First, we applied traditional statistical approaches to differential expression analysis and their adaptations to assumption violations. Second, since in proteomic studies it is not uncommon to come across data that are nonnormally distributed and/or differently dispersed, we discuss two different ways of dealing with these situations. In section 2.4.2., we describe an approach that we refer to as an indirect method where traditional statistical analysis is conducted on the transformed data. In Section 2.4.3., we describe our direct approach, where general classes of distributions, (generalized gamma, exponential, or Weibull) are directly fitted to the data using a generalized linear model.

2.4.1. Differential Protein Expression. For a given 2-D gel experiment, proteomic differential expression analysis describes the process of conducting multiple hypothesis tests, one for each protein, across all commonly expressed proteins. In the traditional two-sample t-test, any protein resulting in a p-value that is less than a pre-specified α (i.e., 0.05) is considered significant and that protein is deemed differentially expressed. This approach must be implemented with caution, because the error rate is fixed only for one specific test and if more than one hypothesis/protein is tested then the error rate accumulates across all tests.

Since many experiments involve the comparison of two treatment groups and since the approach can be easily generalized to more than two-groups, we focus our attention on the two-population comparison. If there are n_1 and n_2 gels processed in groups 1 and 2, respectively, and the populations are assumed to have approximately equal variances, then the two-sample t-test involves the computation of the following test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(1/n_1 + 1/n_2)}}, \quad s_p^2 = [((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)] \quad (1)$$

where $\bar{x}$, $\bar{x}_2$, s^2, and s_2^2 are the sample means and variances from each sample, respectively. In the two-sample t-test, if the normality assumption is reasonable but the common variance assumption is violated then eq 1 may not be valid. However, approximate t-tests are available to test for differences between the two means. Cochran and Cox proposed an approximate t-test, but the degrees of freedom were undefined when the sample sizes were unequal and the test was quite conservative.[27,28] Satterthwaite's approximation for the degrees of freedom can be used for the approximate t-test in these cases, but the test given below, still remains conservative.[29,30]

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_1^2/n_2}}, \; df = (w_1 + w_2)^2/(w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1)), w_1 = s_1^2/n_1, w_2 = s_2^2/n_2)^{29-30} \quad (2)$$

Regardless of the test used (eqs 1 or 2) there is a question of efficiency since one usually tests the equal variance assumption before deciding between the t-test (equal variances) or the approximate t-test (unequal variances). Nonparametric approaches do not provide much relief since they typically assume symmetrically distributed populations with common variance or similar shapes across groups.

Since a 2-D gel experiment involves several hundred hypothesis tests on unknown proteins, it is impossible to know whether the equal variance assumption is true for all proteins. On the basis of the sample expression between two groups, one can test whether the underlying populations have the same spread, dispersion or variance ($\sigma_1^2 = \sigma_2^2$versus $\sigma_1^2 \neq \sigma_2^2$) using the Folded Form F-test (eq 3)

$$F = \max(s_1^2, s_2^2)/\min(s_1^2, s_2^2) \quad (3)$$

which, under the normal distribution assumption, has an F_{n_1-1,n_2-1} distribution. Although most researchers conduct such a test to determine whether the two-sample t-test based on equal variance is valid, we propose that proteins that have significantly different variances between groups may well be of biological significance. That is, if a protein has significantly different variances between groups, then it is included in the list of significant proteins whether there are significant mean differences in expression.

2.4.2. Transformation Approach. In genomic and proteomic studies statistical analyses is often conducted on the log-transformed data across all genes and proteins. In many cases, this approach results in more symmetrically distributed data and/or dampens the effect of nonconstant variance at high levels and outliers. However, Rocke, and Durbin provided evidence that for low expressing genes or proteins, this transformation can make matters worse.[31,32] Accordingly, much literature has been dedicated to more generalized transformations such as Box-Cox[33] and Generalized-log transformations that serve as an alternative to blind application of a single transformation.[32,34] Specifically, if y represents the expression value for a particular protein or gene, then the simple Box-Cox transformation is of the form $z = (y^\lambda - 1)/\lambda$ if $\lambda \neq 0$ and $z = \log(y)$ if $\lambda = 0$.[33] This class includes most of the common transformations, including the log-transform and various power and inverse power transforms. The underlying goal in using a generalized transformation is that the resulting data will be more in line with the model assumptions and therefore produce more robust analyses of the data.[31,32,34–38] The generalized class of transformations are appealing because they are very flexible. They include a form of the simple log-transform as a special case, and the appropriate transformation can be estimated using maximum likelihood approaches.[35] The TRANSREG procedure in SAS offers the maximum likelihood approach in fitting the optimal Box-Cox transformations to data taken from Draper and Smith.[39] The model fitting feature allows one to optimize and/or customize the transformation for each individual protein or gene rather than doing a single log-transform across all proteins or genes.

2.4.3. Generalized Model (GM). The generalized model more directly addresses the problems discussed above by providing a unified theoretical and conceptual framework for analyzing protein differential expression across each protein spot. Generalized models assume the response variable (expression) is not necessarily normally distributed and the underlying distributions may not have constant variances between groups or across levels of the predictor variables.[40] In many cases other than the normal distribution, the populations may have a mathematical dependency (link function) between the variance and the mean of the populations. The GENMOD procedure in SAS provides Newton–Raphson algorithm (ridge-stabilized) to maximize the log-likelihood function in estimation and testing of parameters in the model for a broad collection of models, including the normal, inverse-Gaussian, gamma, negative binomial, and Poisson distributions.

Therefore, we propose a method for 2-D gel analysis whereby the underlying distribution is modeled directly as a generalized-gamma distribution, which has the Weibull, exponential, gamma, and log-normal as special cases. Each of these special distributions has a relationship with a log-location-scale family of distributions. For example, taking the log-transformation of Weibull, gamma, and log-normal data leads to the extreme value distribution, the log-gamma distribution and the normal distribution, respectively. Each of these distributions is a special case of the generalized log-gamma distribution.[40] Therefore, under the right conditions, fitting the generalized gamma or the generalized log-gamma distribution to data leads to distributions approximating the true underlying distributions individually and perhaps more accurate statistical contrasts between treatment groups. Inference can then be made about the location, shape and scale of the distribution without having prior knowledge of the specific positive support distributions across all proteins within the given populations or treatment groups. Therefore, when the goal is discovery of proteins, we propose a method where the generalized gamma distribution is fit to each specific commonly expressed protein within populations and tested for significant differences across the populations. The new list of proteins is then compared and contrasted to those found as worthy of follow-up analysis through other more traditional methods, including tests on mean and variance differences.

Our GM method is expressed as follows: Y denotes the expression for a particular spot on a 2-D gel, and Y has a generalized gamma distribution if its distribution function is of the form

$$f(y) = \frac{|\delta|}{y \cdot \Gamma(\delta^{-2})}(\delta^{-2}y^{\delta})^{\delta-2} \exp(-\delta^{-2}y^{\delta}), y > 0 \qquad (4)$$

where $\Gamma(\cdot)$ is defined as the gamma function. Taking the log-transform of a generalized-gamma random variable, $z = \log(y)$ results in the location-scale family called the generalized-log-gamma distribution, given in its standard form as follows:

$$f(z) = \frac{|\delta|}{\Gamma(\delta^{-2})}(\delta^{-2} \exp(\delta \cdot z))^{\delta-2} \exp(-\delta^{-2} \exp(-\delta \cdot z)),$$
$$z,\delta \in (-\infty,\infty) \qquad (5)$$

The parameter δ is referred to as the shape parameter. If $\delta = 1$, then the log-generalized gamma becomes the extreme value distribution and the corresponding generalized gamma becomes the Weibull distribution. If $\delta = 0$, then the log-generalized gamma becomes the normal distribution and the corresponding generalized gamma becomes the log-normal distribution. Regression analysis based on these models can be done by using the LIFEREG, NLP, or NLIN procedures in SAS. The typical approach is to log-transform the data first, and then fit the generalized log-gamma distribution separately to each of the protein expression variables, which is equivalent to fitting the corresponding generalized gamma to the raw data. The density in eq 5 is expressed in standard form (just as a normal distribution with mean zero and unit variance is the standard form of the normal family). As Lawless pointed out, the generalized log-gamma (GLG) distribution is a location-scale family, just like the normal family, and these parameters are introduced into the density by letting $z = (u - \mu)/\sigma$, and u becomes a GLG (μ,σ,δ), where μ, σ, and δ represent the location, scale and shape parameters, respectively.[41]

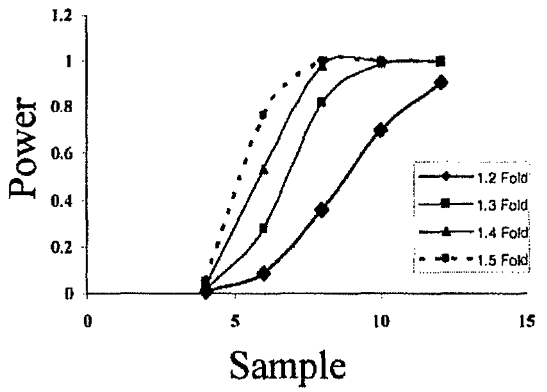

Figure 1. Power analysis versus sample size. This graph illustrates how power and sample size are related with respect to detection of fold change in protein expression.

The location can be expressed in terms of linear regression model on the log-transformed data. Initial estimates of regression parameters are obtained by doing ordinary least-squares regression on the log-transformed data, which are then used to get more precise maximum-likelihood estimators (MLE) using some numerical method such as ridge-stabilized Newton–Raphson algorithm. Differences in location between populations can then be tested directly using a χ-square test (Wald test). The GENMOD or NLP procedures could be used to generalize this approach for simultaneously testing for differences in location (mean/median), scale (variances/standard deviation) and shape, but for illustrative purposes in this paper we focus on tests for differences in location in models in which the mean and variance are mathematically related.

Results

3.1. Power Analysis/Sample Size Determination. In our first pilot study, five replicate 2-D gels from a single uterine sample were used to examine reproducibility. The results showed that the total number of protein spots per gel were reasonably similar. However, when we looked only at common spots among the gels, we found that as the multiplicity of gels increased there was a significant decrease in number of common spots (unpublished data). This is consistent with earlier findings reported by Voss and Haberl.[7]

A second pilot study using uteri from 8 control- and 8 genistein-treated rats also showed that as the multiplicity of gels increased there was a significant decrease in number of common spots. Since this decrease in matched spots was at similar rates between groups, it indicated a lack of sample handling bias. The pooled control estimate of standard deviation in normalized peak intensities was used to determine that a sample size of 8 animals per treatment group would be sufficient to detect a 1.5-fold-change between the two groups. This change was detected with over 99% power, based on a two-sample t-test with an experiment-wise level of significance of $p < 0.05$, with adjustments for multiple testing. Figure 1, displays the power curves for the detection of four different fold changes (1.2, 1.3, 1.4, and 1.5). The power is defined as the probability of detecting the specified fold-change and is displayed over sample sizes ranging between 4 and 12. While a sample size of 8 gives 99% power to detect a 1.5-fold-change, this power drops to 82% to detect a 1.3-fold-change. A sample size of 6 only gives 28% power to detect a 1.3-fold-change.

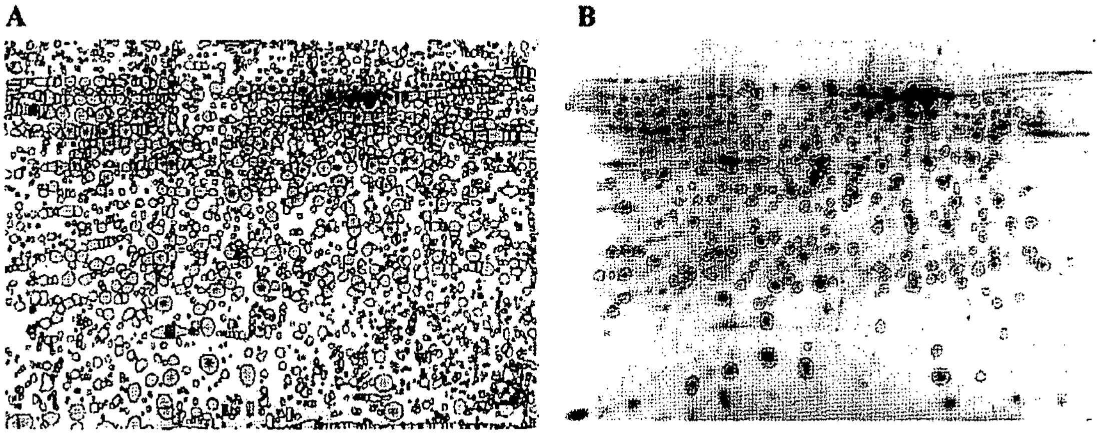

Figure 2. 2-D gel profile (A) A display of unsupervised spot detection results. (B) A display of those spots common to all gels in the experiment.

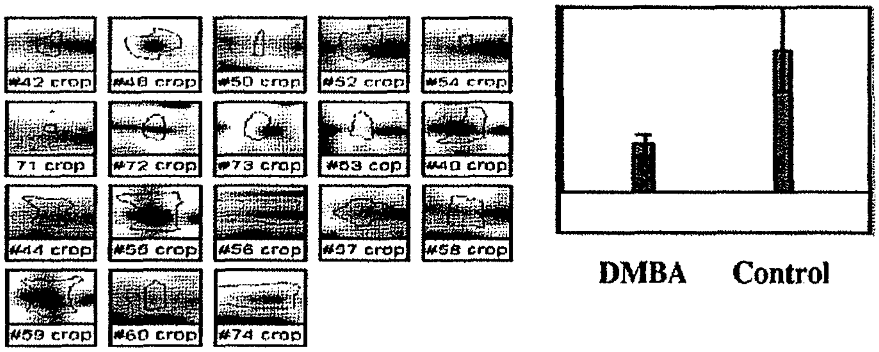

Figure 3. Supervised spot evaluation. Initial evaluation of common spots is based on the p-value from a two-sample t-test. The graph shows the mean value (±SEM) for the normalized value. Spots are ranked according to p-values and all spots with a $p < 0.05$ are subjected to visual inspection to verify consistency of spot parameters. This figure shows that while the t-test on the normalized value was significant ($p = 0.0431$), there is inconsistency in spot detection.

3.2. Data Quality and Processing. Results of unsupervised matching and spot detection (Figure 2A) demonstrate the need for a directed process of image cleanup before evaluation. After initial matching we focused only on those spots found in all gels (Figure 2B). Common spots with significant p-values ($p <$ 0.05, t-test) were subjected to visual verification to ensure both accuracy of matching and consistency of spot boundary (Figure 3). This early analysis is critical to prevent improper data interpretation. Once several new landmarks have been established the matching program and inspection process is rerun. This iterative process greatly increases the efficiency of subsequent evaluations by providing well matched data points for the more robust statistical procedures.

3.3. Statistical Analysis of Two Experimental Groups. Our primary data set was generated using 18 gels representing unique mammary gland samples in each of two treatment groups (10 control and 8 DMBA treated rats). Analysis of all 18 gels yielded 247 spots that were present in every gel. These 247 common spots were subjected to statistical differential expression analysis. Evaluation of the data using only the t-test on the untransformed data found 13 spots to be significantly different between the 2 groups ($p < 0.05$) (Figure 4A). Testing of the log-transformed data yielded a total of 15 spots to be significantly different ($p < 0.05$) (Figure 4B). GM calculations added an additional 11 spots for a total of 26 spots that were significantly different ($p < 0.05$) (Figure 4C). Using a 0.05 level of significance, the estimated FDR was 0.20. Therefore, we expected 5 of the 26 spots found using the GM to be false positives.

3.3.1. Generalized Models. An advantage of the GM procedure is that it allows for the mean and variance to be linked and vary simultaneously between groups. Individual data plots for three spots where the p-values differed for the t-test, log-normalized and GM are presented in Figure 5. For each protein spot the individual data points of mammary glands of control- and DMBA-treated animals are graphed to show the variation for the log-normalized data. For those instances where the norm and log-normalized data are not significantly different we assume that the mean values are similar. Graphs in Figure 5A,B demonstrate that while the means are similar, the underlying variation of expression is different. Therefore, using the GM we model this variation and determine the spots to be significantly different ($p < 0.05$).

3.3.2. Tests on Equal Variances (Folded Form F-test). Figure 6 illustrates that using just the Folded Form F-test (testing only on variance) we find 33 unique proteins not captured by any of the other tests. Finally, we see that there is an overlap of only 3 spots identified as being significant using all testing procedures. The field graph in Figure 7 illustrates all 247 protein spots that were evaluated. This graph reveals the overlap of significantly evaluated spots. Each significant spot's location is based on either differences only in the variance as a function

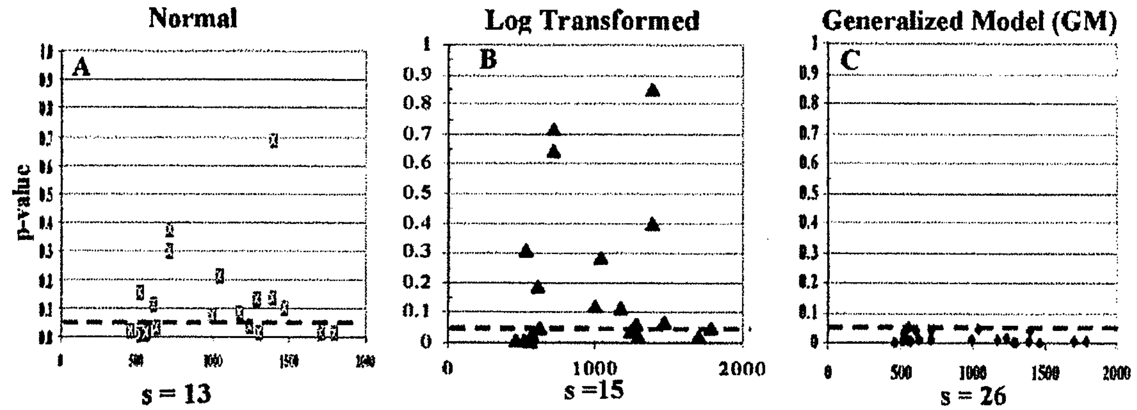

Figure 4. Comparison of traditional and GM testing procedures. The number of significant spots for each testing procedure is represented below the dashed line ($p < 0.05$). (A) The t-test applied to the normalized data found 13 spots to be significantly different. (B) Means testing on the logtransformed data found 15 spots differentially expressed and (C) results using the GM captured 26 spots as differentially regulated.

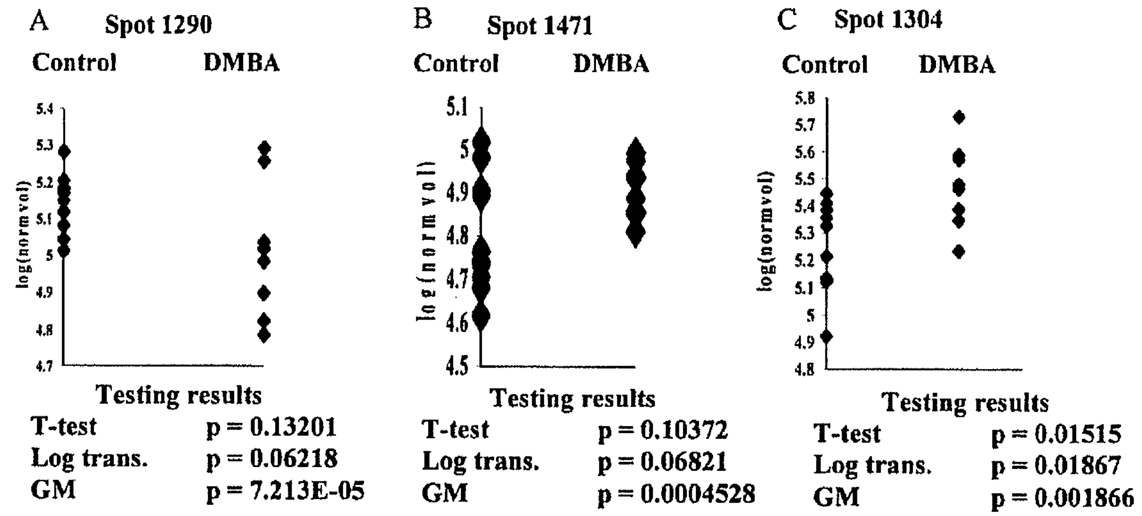

Figure 5. Consideration of variance. Each graph displays the log-transformed data for an individualsample in either group (Control or DMBA) (A) Results of the two-sample t-test on the normal or logtransformed data for spot 1290 are not significant ($p > 0.05$). However, results of the GM show a highly significant ($p = 7.213 \times 10^{-05}$) difference in variance between the control and DMBA groups. (B) For Spot 1471 results of the two-sample t-test on the normal or log-transformed data are not significant ($p > 0.05$). However, the GM found a highly significant ($p = 0.000\,452\,8$) difference in variance between the control and DMBA groups. (C) For spot 1304, the results of the two sample t-test were significant and with log transformation ($p < 0.05$), as well the GM was significant ($p = 0.001\,866$).

of the same mean, or variance in the absences of similar means. Finally, spots uniquely found significant using the GM procedure are distinguished in the broad field.

4. Discussion

Proteomics and genomics fall under the general heading of systems biology. Systems biology focuses on the interaction of all molecular components including: DNA, RNA, proteins, protein interactions, biomodules, cells, tissues, etc., with each of these components having their own individual elements (e.g., specific gene methylation or protein post-translational modifications). A systems level view is necessary to understand the complex dynamics that underlie the physiology in both the normal and diseased states. Systems biology is characterized by a synergistic integration of theory, computation, and experiment.[42]

Advances in recent technology make possible the large-scale application of proteomics for biomarker discovery in cancer models and the exploration of mechanisms of action of drugs. These advances result in the ability to readily run reproducible 2-D gels for protein separation and obtain protein identification using mass spectrometry techniques, such as MALDI-TOF. Software programs, such as Progenesis, have been developed that aid the researcher in evaluating changes in protein expression profiles among groups and between samples. However, these programs lack substantial statistical analysis tools to help researchers determine the most important and persistent changes throughout the experiment. Without adequate means of analysis the researcher is left to generate a long list of proteins for identification, and then is required to use a hit-or miss strategy for further analysis.

The 2-D gel cleanup/spot review and evaluation cycle has long been considered the bottleneck of 2-D gel experiments. This has resulted from over reliance on the unsupervised matching and spot evaluation by the software followed by an unscripted procedure for cleanup by the end-user. Therefore,

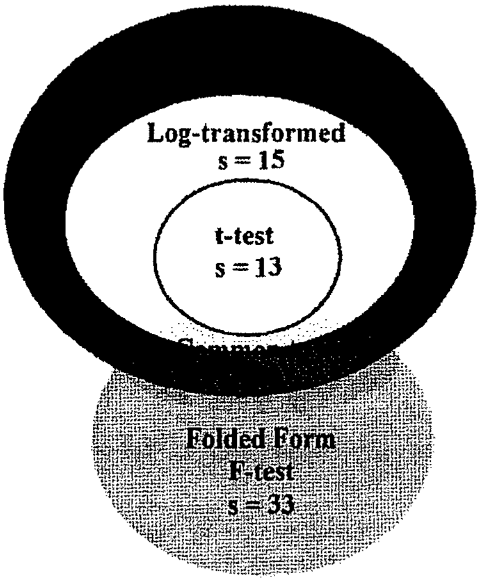

Figure 6. Nested collection of significant spots. The GM procedure found the same spots as the traditional t-test as well as those found from testing on the log-transformed data. The number of unique spots using the Folded Form F-test (s = 33) are demonstrated by the green circle. Three spots were found to be significant regardless of testing method.

we have developed a method that greatly increases the speed of this process by providing guidance and direction. Through multiple trials we have determined that statistical analyses are best conducted only on the common spots.

Our current research focuses on finding biochemical events that indicate the earliest stages of breast cancer development. Using an animal model of carcinogenesis, we developed our evaluation of markers along a known timeline of tumor development. The DMBA model we chose has been demonstrated to result in 100% mammary tumor incidence. In general, we have seen that DMBA treatment to 50 day old rats results in palpable tumor development when the animals are 100–120 days old; therefore our choice to evaluate mammary glands at 75 days of age (25 days after DMBA administration) represents a very early state of carcinogenesis. Pathological examination of these animals showed no lesion formation in the DMBA treated animals at day 75. Given that cancer is a disease process with a long developmental period we acknowledge that the earliest stages of carcinogenesis are likely marked by subtle alterations in protein expression. These low expression differences are one reason that we have emphasized power analysis to provide information about our lower limits of detection in 2-D gel experiments.

Power analysis is a standard method to determine a level of sensitivity for value change (such as spot volume fold change) as a function of the sample size. In any biomedical experiment, the number of experimental units (sample size) should be selected to maximize the probability (power) of detecting a predetermined significant difference between two or more treatments (i.e., protein fold change). By addressing the issue of sensitivity from the beginning, this knowledge can be applied to help determine if the changes in expression of a particular protein make logical sense for the given experimental design/biology. While replication studies for power determination can be costly, establishment of statistically relevant data will lead to reduced end-cost. For our biological model, the result of power and sample size determination established our ability to confidently identify those spots that differed in mean expression by 1.5-fold or greater with a reasonable number of biological replicates. However, results of traditional expression evaluation, t-test and log transformed data, only identified a finite number of significant spots (s = 13 and s = 16, respectively). In fact, finding 13 to 16 spots represents only 5–6% of the total evaluated protein spots (s = 247). This low value, while technically accurate, represents a level of finding

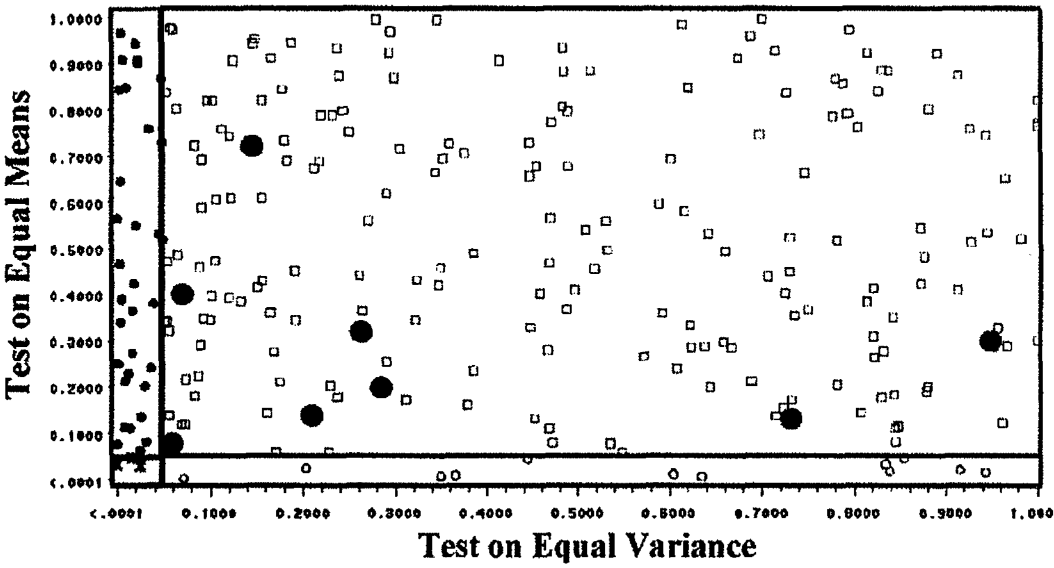

Figure 7. Field graph of commonly expressed spots. The vertical axis represents the p-values from a two-sample t-test conducted on the log-transformed data (shaded area represents $p < 0.05$). The horizontal axis represents the p-values resulting from a test on equal variances among the groups (shaded area represents $p < 0.05$). Red circles depict those p-values based on the generalized linear model (GM) that were significant.

not unlikely to be based on chance. Therefore, we needed to design a more robust approach to evaluate our data.

As mentioned in the Introduction, the application of the GM provides additional information on the distribution of the individual data points for a particular spot. By applying the concept of NoR to our evaluation we saw alteration in the variance levels, either tighter regulation or dysregulation for some of the proteins examined, while the mean appears similar (see Figure 5). Spread of variation shows the natural characteristic of the model to allow for wide fluctuations in the normal circumstance, or the inverse that certain proteins require strict control to maintain adequate cellular function. By evaluating the changes in variance of expression we gain insight into a level of control that may be involved in the promotion of carcinogenesis. Since the result of 2-D gels is to look at a broad spectrum of proteins, we may be able to establish patterns of variance alteration and determine if proteins that undergo positive or negative shifts in expression are functionally related to one another or the disease process.

The changes of any particular protein over the course of tumor development will itself alter as, in the case of mammary cancer, the underlying cell population changes.[14] Traditionally, tumorogenesis is measured as a mean time to tumor development, hence we have to use multiple animals/group to get mean to first, second, etc. tumor/rat since the individual animal's response is different. Furthermore, it is the fundamental effect of treatments such as cancer promoters or chemopreventive agents to alter the time of tumor development. However, all of these measures ignore the individual response or the general group response unless the mean levels are significantly different. Ultimately, we appreciate that underlying alterations involved in the long term process of carcinogenesis will likely be found in subtle, yet persistent, changes in cellular signaling.

It is well recognized that the value for an individual spot on a 2-D gel does not necessarily represent an absolute measurement for the concentration of a protein. For this reason, we acknowledge that there is some inherent weakness in performing exhaustive evaluations of spots from a statistical standpoint. It is our assumption that investigators are willing to make certain tradeoffs in data quality vs time and future evaluation. That is to say, any mass-spectrometry based protein identification is going to require more stringent confirmation procedures, such as immuno techniques. In turn, these techniques will allow for a more quantitative assessment of changes in protein concentration. It is our intent to provide more information about the general qualities of the information that the 2-D gel is providing and to help guide the researcher in the decision making process with respect to which spots should be evaluated first. Therefore, displays such as Figure 7 provide all of the information with respect to what model resulted in a spot being found significant. This system alleviates the production of laundry-lists of proteins and allow for directed and focused studies of particular proteins/pathways that are involved in the condition under study. Therefore, our future experiments will be designed to more accurately capture data related to the temporal changes we have observed to better establish the role of identified proteins.

In summary, we have described a reproducible and statistical approach to the use of 2-D gels for identification of biomarkers that may be related to the carcinogenesis of DMBA in the rat mammary gland. These methods lend well to the discovery of novel new proteins and identification of key signaling pathways involved in cancer causation. Our statistical approach involves empirical determination of the number of gels required to ensure statistical power for appropriate evaluation. In general, the approach we used results in quickly identifying those proteins that meet a realistic and significant change, but is also broad enough to allow the unique modeling approach of the GM. The approach that we have outlined is what we consider to be discovery proteomics. Only when we have mass spectrometry data for identification do we consider this as our preliminary data, not as conformational or primary data. Experiments can then be designed to evaluate the validity of identifications including the previous mention of more specific techniques of quantification.

Acknowledgment. C.R. was supported by stipends from DOD BC 17-00-1-0119 and the Susan G. Komen Foundation DISS0201242. This research was funded by Susan G. Komen Foundation DISS0201242, NIH 1 R01 CA61742-07, NIH 1 UO1 ES012771-01, and DOD BC 021371.

References

(1) O'Farrell, P. Z.; Goodman, H. M. Resolution of simian virus 40 proteins in whole cell extracts by two- dimensional electrophoresis: heterogeneity of the major caspid protein. *Cell* **1976**, *9*, 289–298.

(2) Illsley, N. P.; Lamartiniere, C. A.; Lucier, G. W. Analysis of the sex specific changes in rat hepatic cytosol protein patterns using two-dimensional gel electrophoresis. *J. Appl. Biochem.* **1979**, *1*, 385–395.

(3) Arnott, D.; O'Connell, K. L.; King, K. L.; Stults, J. T. An integrated approach to proteome analysis: identification of proteins associated with cardiac hypertrophy. *Anal. Biochem.* **1998**, *258*, 1–18.

(4) Lewis, T. S.; Hunt. J. B.; Aveline, L. D.; Jonscher, K. R.; Louie, D. F.; Yeh, J. M.; Nȧhreini, T. S.; Resing, K. A.; Ahn, A. G. Identification of novel MAP kinase pathway signaling targets by functional proteomics and mass spectrometry. *Mol. Cell.* **2000**, *6*, 1343–1354.

(5) Hondermarck, H.; Vercoutter-Edouart, A. S.; Révillion, F.; Lemoine, J.; El-Yazidi-Belkoura, I.; Nurcombe, V.; Peyrat, J.-P. Proteomics of breast cancer for marker discovery and signal pathway profiling. *Proteomics* **2001**, *1*, 1216–1232.

(6) Taylor, C. F.; Paton, N. W.; Garwood, K. L.; Kirby, P. D.; Stead, D. A.; Yin, Z.; Deutsch, E. W.; Selway, L.; Walker, J.; Riba-Garcia, I.; Mohammed, S.; Deery, M. J.; Howard, J. A.; Dunkley, T.; Aebersold, R.; Kell, D. B.; Lilley, K. S.; Roepstorff, P.; Yates, J. R. 3rd; Brass, A.; Brown, A. J.; Cash, P.; Gaskell, S. J.; Hubbard, S. J.; Oliver, S. G. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.* **2003**, *21*, 247–254.

(7) Voss, T.; Haberl, P. Observations on the reproducibility and matching efficiency of two-dimensional electrophoresis gels: consequences for comprehensive data analysis. *Electrophoresis* **2000**, *21*, 3345–3350.

(8) Nishihara, J. C.; Champion, K. M. Quantitative evaluation of proteins in one- and two-dimensional polyacrylamide gels using a fluorescent stain. *Electrophoresis* **2002**, *23*, 2203–2215.

(9) Maurer M. H.; Feldmann, R. E.; Bromme, J. O.; Kalenka, A. Comparison of statistical approaches for the analysis of proteome expression data of differentiating neural stem cells. *J Proteome Res.* **2005**, *4*, 96–100.

(10) Chang, J.; Van Remmen, H.; Ward, W. F.; Regnier F. E.; Richardson, A.; Cornell, J. Processing of data generated by 2-dimensional gel electrophoresis for statistical analysis: missing data, normalization, and statistics. *J. Proteome Res.* **2004**, *3*, 1210–1218.

(11) Gustafsson, J. S.; Ceasar, R.; Glasbey, C. A.; Blomberg, A.; Rudemo, M. Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data. *Proteomics* **2004**, *4*, 3791–3799.

(12) Woltereck, R. Weitere experimentelle Untersuchungen über Artveränderung, speziell über das Wesen quantitativer Artunterschiede bei Daphnien. *Verh. Deutsch. Zool. Gesellsch.* **1909**, *19*, 110–173.

(13) Sarkar, S.; Fuller, T. Generalized norms of reaction for ecological developmental biology. *Evol. Dev.* **2003**, *5*, 106–15.

(14) Russo, J.; Wilgus, G.; Russo, I. H. Susceptibility of the mammary gland to carcinogenesis: I Differentiation of the mammary gland as determinant of tumor incidence and type of lesion. *Am. J. Pathol.* **1979**, *96*, 721–736.

(15) Lamartiniere, C. A.; Cotroneo, M. S.; Fritz, W. A.; Wang, J.; Mentor-Marcel, R.; Elgavish, A. Genistein chemoprevention: timing and mechanisms of action in murine mammary and prostate. *J. Nutr.* **2002**, *132*, 552S–558S.

(16) Muller, P.; Parmigiani, G.; Christian, R.; Rousseau, J. Optimal sample size for multiple testing: The case of gene expression microarrays. *J American Stat Soc.* **2004**, *99*, 990–1001.

(17) Simon, R. M.; Dobbin, K. Experimental design of DNA microarray experiments. *Biotechniques Suppl.* 2003, 16–21.

(18) Yang, Y. H.; Speed, T. Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* **2002**, *3*, 579–588.

(19) Molloy, M. P.; Brzezinski, E. E.; Hang, J.; McDowell, M. T.; VanBogelen, R. A. Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics* **2003**, *3*, 1912–1919.

(20) Asirvatham, V. S.; Watson, B. S.; Sumner, L. W. Analytical and biological variances associated with proteomic studies of Medicago truncatula by two-dimensional polyacrylamide gel electrophoresis. *Proteomics* **2002**, *2*, 960–968.

(21) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., Series B* **1995**, *85*, 289–300.

(22) Allison, D. B.; Gadbury, G. L.; Heo, M.; Fernández, J. R.; Lee, C.-K.; Prolla, T. A.; Weindruch, R. A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat., Data Anal.* **2002**, *39*, 1–20.

(23) Storey, J. D. A direct approach to false discovery rates. *J. Royal Stat. Soc. Ser. B* **2002**, *64*, 479–498.

(24) Storey, J. D. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Stat.* **2003**, *31*, 2013–2035.

(25) Liao, J. G.; Lin, Y.; Selvanayagam, Z. E.; Shih, W. J. A mixture model for estimating the local false discovery rate. *Bioinformatics* **2004**, *20*, 2694–2701.

(26) Fountaoulakis, M, Schull, E, Hardmeier, R, Bernt, P and Lubec, G. Rat brain proteins: Two-dimensional protein database and variations in the expression level. *Electrophoresis* **1999**, *20*, 3572–3579.

(27) Cochran, W. G.; Cox, G. M. *Experimental Designs*; John Wiley & Sons: Inc.: New York, 1950.

(28) Lee, A. F. S.; Gurland, J. Size and power of tests for equality of means of two normal populations with unequal variances. *J. Am. Stat. Assoc.* **1975**, *70*, 933–941.

(29) Satterthwaite, F. W. An approximate distribution of estimates of variance components. *Biometrics Bull.* **1946**, *2*, 110–114.

(30) Freund, R.J.; Littell, R. C.; Spector, P. C. *SAS System for Linear Models*; SAS Institute Inc.: Cary, NC, 1986.

(31) Rocke, D. M.; Durbin, B. A model for measurement error for gene expression arrays. *J. Comput. Biol.* **2001**, *8*, 557–569.

(32) Rocke, D. M.; Durbin, B. Approximate variance-stabilizing transformations for gene-expression microarry data. *Bioinformatics*, 2003, *19*, 966–972.

(33) Box, G. E. P.; Cox, D. R. An Analysis of Transformations. *J. Royal Stat. Soc.* **1964**, *B-26*, 211–252.

(34) Cui X.; Churchill G. A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **2003**, *4*, 210.

(35) Foster; A. M.; Tianm, L.; Wei, L. J. Estimation for the Box-Cox transformation model without assuming parametric error distribution. *J. Am. Stat. Assoc.* **2001**, *96*, 1097–1101.

(36) Geller, S. C.; Gregg, J. P.; Hagerman, P.; Rocke, D. M. Transformations and normalization of oligonucleotide microarray data. *Bioinformatics* 2003, *19*, 1817–1823.

(37) Huber, W.; von Heydebreck, A.; Sultmann, H.; Poustka, A.; Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **2002**, *18*, 96–104.

(38) Cui, X.; Churchill, G. A. Transformations for cDNA microarray data. *Stat. App. Gen. Mol. Biol.* **2003**, *2*, 1.

(39) Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: New York, 1982.

(40) McCullagh, P.; Nelder, J. A. *Generalized Linear Models*; Chapman & Hall/CRC Press: Boca Raton, FL 1998

(41) Lawless, J. F. *Statistical Models and Methods for Lifetime Data*; Wiley Series in Probability and Statistics 2nd ed.; John Wiley & Sons: New York, 2003.

(42) Workshop Report: *Defining the Mandate of Proteomics in the Post-Genomics Era*; National Research Council: The National Academies Press: Washington, D. C., 2002.

PR0501261